02 117
RSW

日　本　国　特　許　庁
JAPAN PATENT OFFICE

別紙添付の書類に記載されている事項は下記の出願書類に記載されている事項と同一であることを証明する。

This is to certify that the annexed is a true copy of the following application as filed with this Office

出願年月日
Date of Application: ２００２年　７月２６日

出願番号
Application Number: 特願２００２－２１８７４０

[ST.10/C]: [ＪＰ２００２－２１８７４０]

出願人
Applicant(s): 日本アイ・ビー・エム株式会社

２００３年　３月　７日

特許庁長官
Commissioner,
Japan Patent Office 太田信一郎

出証番号　出証特２００３－３０１４３２４

【書類名】　特許願

【整理番号】　JP9020117

【提出日】　平成14年 7月26日

【あて先】　特許庁長官　殿

【国際特許分類】　G06F　9/44

【発明者】

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名】　槌谷　一

【発明者】

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名】　村上　佐枝子

【発明者】

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名】　豊島　浩文

【発明者】

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名】　日高　由布子

【特許出願人】

【識別番号】　390009531

【氏名又は名称】　インターナショナル・ビジネス・マシーンズ・コーポレーション

【代理人】

【識別番号】　100086243

【弁理士】

【氏名又は名称】　坂口　博

【代理人】

【識別番号】　100091568

【弁理士】

【氏名又は名称】　市位　嘉宏

【代理人】

【識別番号】　100108501

【弁理士】

【氏名又は名称】　上野　剛史

【復代理人】

【識別番号】　100104880

【弁理士】

【氏名又は名称】　古部　次郎

【手数料の表示】

【予納台帳番号】　081504

【納付金額】　21,000円

【提出物件の目録】

【物件名】　明細書　1

【物件名】　図面　1

【物件名】　要約書　1

【包括委任状番号】　9706050

【包括委任状番号】　9704733

【包括委任状番号】　0004480

【プルーフの要否】　要

【書類名】　明細書

【発明の名称】　情報収集システム、アプリケーションサーバ、情報収集方法、およびプログラム

【特許請求の範囲】

【請求項１】　複数のデータファイルをネットワークを介して取得する取得手段と、

前記取得手段により取得された複数のデータファイルに対し、所定の切り出しルールと用語の関係記述であるオントロジとを利用して解析する解析手段と、

前記解析手段による解析に基づいて、前記複数のデータファイルから必要な情報を抽出する抽出手段と

を含む情報収集システム。

【請求項２】　前記取得手段は、ユーザの興味に基づくＵＲＬ(Uniform Resource Locators)を用いてＨＴＭＬ(Hypertext Markup Language)で書かれた文書を取得し、

前記解析手段は、前記特定のタグ情報を用いて前記文書を解析することを特徴とする請求項１記載の情報収集システム。

【請求項３】　前記解析手段による解析に用いられる前記所定の切り出しルールは、カタログおよび/または仕様情報を構成する特徴をルール化したものであることを特徴とする請求項１記載の情報収集システム。

【請求項４】　前記解析手段は、用語の異なる複数のデータファイルに対し、オントロジを利用して横断的に内容を解析することを特徴とする請求項１記載の情報収集システム。

【請求項５】　前記抽出手段により抽出された情報を再構築し、当該情報の中から同値関係をまとめてユーザ端末に提供する提供手段と

を更に含む請求項１記載の情報収集システム。

【請求項６】　対象ごとに異なったオントロジを格納するオントロジ格納手段を更に備え、

前記解析手段は、前記オントロジ格納手段から所定のオントロジを読み出して解析を行うことを特徴とする請求項１記載の情報収集システム。

【請求項７】　ユーザの興味に関する情報を受信するユーザ要求受信部と、

前記ユーザ要求受信部より受信した前記情報に基づいて、複数のサイトからＨＴＭＬ文書を取得するＨＴＭＬ取得部と、

前記ユーザ要求受信部より受信した前記情報に基づいてオントロジを読み出し、語彙情報を得る語彙情報処理機構と、

前記ＨＴＭＬ取得部から取得した前記ＨＴＭＬ文書に対し、前記語彙情報処理機構から提供される前記語彙情報に基づき、当該ＨＴＭＬ文書のタグを頼りに抽出データオブジェクトを取り出す抽出位置情報特定部と

を含むアプリケーションサーバ。

【請求項８】　前記ＨＴＭＬ取得部から取得した前記ＨＴＭＬ文書に対して切り出し処理を施すための切り出しルールを提供する切り出しルール処理機構を更に備え、

前記抽出位置情報特定部は、前記切り出しルール処理機構から提供される切り出しルールに基づいて抽出データオブジェクトを取り出すことを特徴とする請求項７記載のアプリケーションサーバ。

【請求項９】　公理ルールに基づいて推論演算を実行する推論処理機構を更に備え、

前記抽出位置情報特定部は、前記推論処理機構にて実行される推論演算に基づいて抽出データオブジェクトを取り出すことを特徴とする請求項７記載のアプリケーションサーバ。

【請求項１０】　前記抽出位置情報特定部によって取り出された複数の抽出データオブジェクトに対して合算処理を施す情報整理集約部と、

前記情報整理集約部による合算処理の結果に基づいてテーブルおよび/またはリストを生成する合算結果オブジェクト生成部と、

前記合算結果オブジェクト生成部により生成された合算結果オブジェクトを送信するユーザ要求送信部とを更に備えたことを特徴とする請求項７記載のアプリケーションサーバ。

【請求項１１】　ネットワークに接続されたコンピュータにおいて、

複数のデータファイルをネットワークを介して取得するステップと、

取得された前記複数のデータファイルに対し、所定の切り出しルールと用語の関係記述であるオントロジとを利用して解析するステップと、

解析された前記複数のデータファイルから有用な情報を抽出するステップと、

抽出された前記有用な情報をユーザにとって利便性のよい形で再構築するステップと

を含む情報収集方法。

【請求項１２】　ネットワークを介して取得されるＨＴＭＬの文書からTableタグまたはListタグに基づいて情報を抽出するステップを更に含む請求項１１記載の情報収集方法。

【請求項１３】　前記解析するステップは、カタログおよび/または仕様情報を構成する特徴をルール化した前記切り出しルールを用いてテーブルの位置決めを行うステップと、位置決めされたテーブルの見出しがユーザにより指定されたカテゴリで一般に使用されている語彙であるか否かの語彙情報を定義する前記オントロジによって語彙のゆれを平準化するステップと、を含むことを特徴とする請求項１１記載の情報収集方法。

【請求項１４】　インターネットに接続されたコンピュータにおいて、

ユーザの興味に関する情報を受信するステップと、

前記ユーザの興味に基づき、インターネットを介して複数の文書を取得するステップと、

格納されている複数のオントロジから、前記ユーザの興味に基づいて特定のオントロジを選定するステップと、

選定された前記特定のオントロジを用いて、取得された前記複数の文書に対して横断的に内容を解析し、有用な情報を抽出するステップと、

を含む情報収集方法。

【請求項１５】　前記ユーザの興味に関する情報は、対象ＵＲＬとオントロジ種別に関する情報であり、

前記複数の文書を取得するステップは、前記対象ＵＲＬに基づいてＨＴＭＬ文書を取得し、当該ＨＴＭＬ文書からテーブル部分またはリスト部分を抜き出すことを特徴とする請求項１４記載の情報収集方法。

【請求項１６】　ネットワークに接続されたコンピュータにおいて、

関連性のある内容に対して各々異なった語彙にて表現された情報を含む複数のＷｅｂページを取得し、

取得した前記複数のＷｅｂページからTableタグまたはListタグに基づいて情報を抜き出し、

抜き出した情報に対して、語彙と語彙との関連付けを示すオントロジに基づき当該複数のＷｅｂページの異なる語彙に対して横断的に情報を解析し、

解析された情報を合算し、

合算結果をユーザ端末に対して送信することを特徴とする情報収集方法。

【請求項１７】　前記合算は、各Ｗｅｂページで異なる語彙に対して、上位下位概念の処理および/または関係処理を施して項目のマッチングを行うことを特徴とする請求項１６記載の情報収集方法。

【請求項１８】　コンピュータに、

複数のデータファイルをネットワークを介して取得する機能と、

取得された前記複数のデータファイルに対し、所定の切り出しルールと用語の関係記述であるオントロジとを利用して解析する機能と、

解析された前記複数のデータファイルから有用な情報を抽出する機能と、

抽出された前記有用な情報をユーザにとって利便性のよい形で再構築する機能と

を実現させるプログラム。

【請求項１９】　前記解析する機能は、所定の公理ルールに基づいて推論演算を実行することを特徴とする請求項１８記載のプログラム。

【請求項２０】　前記再構築する機能は、関連性のある語彙と値について同値関係を処理し、更に新たな関係を挿入して情報を再構築することを特徴とする請求項１８記載のプログラム。

【請求項２１】　コンピュータに、

ユーザの興味に関する情報に基づき、インターネットを介して複数の文書を取得する機能と、

格納されている複数のオントロジから、前記ユーザの興味に基づいて特定のオ

ントロジを選定する機能と、

選定された前記特定のオントロジを用いて、取得された前記複数の文書に対して横断的に内容を解析する機能と、

を実現させるプログラム。

【請求項２２】　コンピュータに、

関連性のある内容に対して各々異なった語彙にて表現された情報を含む複数のＷｅｂページを取得する機能と、

取得した前記複数のＷｅｂページからTableタグまたはListタグに基づいて情報を抜き出す機能と、

抜き出した情報に対して、語彙と語彙との関連付けを示すオントロジに基づき当該複数のＷｅｂページの異なる語彙に対して横断的に情報を解析する機能と、

解析された情報を合算する機能と、

を実現させるプログラム。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】

本発明は、情報を収集・整理する情報収集システム等に係り、より詳しくは、例えばＷｅｂ上に公開されている様々な領域のカタログ情報等につき、所定の抽出ルールに基づき、例えば同類項目を合算して表示等を行う情報収集システム等に関する。

【０００２】

【従来の技術】

近年、インターネット利用の普及に伴い、例えば車やパーソナルコンピュータ（ＰＣ）、不動産、金融関係等の情報をユーザが必要とする際、各サイトからＷｅｂ（ワールド・ワイド・ウェブ：WWW）を通じてＷｅｂコンテンツの配信を受けることが一般的に行われている。これらの情報を必要とする際に、ユーザは、自動車会社のホームページ（ＨＰ）やコンピュータ会社のＨＰ等からカタログ情報等を取得し、取得したこれらのカタログ情報等を比較検討して商品購入を決定している。

【０００３】

ここで、これらのカタログ情報等は、各種情報が項目別に分類されたテーブル形式を用いてユーザに提供されており、それ自身としては、ユーザに対して見易い形式、見易い内容となるように工夫されている。しかしながら、これらの情報は、各社の独自な基準で作成されており、ユーザによる比較検討が非常に難しい。例えば、ＰＣのカタログを例に挙げると、例えば、Ａ社では「ＣＰＵ」という文言が用いられ、Ｂ社では「プロセッサ」という文言が用いられており、同様な意味について異なった文言が用いられている場合がある。また、ノートブック型ＰＣでは、例えば、Ａ社では「バッテリ重量」と「本体重量」という文言で表記され、Ｂ社ではこれらを合わせて「総重量」と表記されている場合など、文言や表記の仕方が異なっている場合もある。

【０００４】

従来では、これらの情報について比較検討する際、ユーザが一つ一つのサイトを開き、手作業で比較することが行われていた。また、自動車会社などでは、各車両のデータについて、各メーカからの公開情報(カタログ・リリース等)から担当者が抜粋し、各装備類の名称等について、その会社の名称に統一、分類して表記されたものをユーザに提供している例もある。

【０００５】

【発明が解決しようとする課題】

しかしながら、従来、これらの作業は、上述のように人間が手動で行っていることから、比較検討に多大な時間を要すると共に、必ずしも正確な検討結果が得られるものではない。また、例えば、自社の名称にて統一して比較結果を提供する場合でも、従来では人間が手動で名称の統一や更新を行う必要があり、ユーザに対してタイムリーな情報提供が困難であった。また、例えば自動車の比較結果を自動車会社が提供する上記場合においても、車種の最新情報等については更新が遅れる場合も多く、最終的な最新情報の確認は、ユーザにより各メーカのＨＰ、カタログ等で行うことを余儀なくされていた。

【０００６】

そこで、インターネット上の複数の情報を機械的に取り出すことが望まれる。

しかし、各サイトから提供されるＷｅｂページは、現在、ほぼＨＴＭＬ形式のみで記述され、記載されているテーブルは、単に、ユーザの見易さだけが念頭に置かれている。そのために、非常に複雑なテーブル構造、複雑なツリー構造となっており、簡単には必要な情報を取り出すことができない。また、これらの情報は、機械的に見て構造化されていない文書と言うことができ、例えば、ページの中でどこに情報があるのか、を機械的に把握することは難しく、更に、同じ概念が違う言葉で表現されており、ユーザが情報を入手した後の機械的な二次処理は困難である。

【０００７】

また、例えば、価格情報提供サイトのように、様々なデータの集計情報を提供するサイトが存在するが、これは所謂Screen Scrapingという方法(各社のＨＴＭＬの構成を作り込みでプログラムすることで、必要な情報を得る方法)で実現されており、情報提供サイトのＨＴＭＬ構造が変わると情報収集ができなくなっていた。そのために、自前のデータベースに人手を介して情報を入力するものが大半であった。

【０００８】

また、例えば、テレビ番組を逃さずチェックしてくれるソフトツールも存在する。このソフトツールでは、ユーザが類義語を定義し、各社テレビガイドのＷｅｂページからテレビ番組の情報を取得し、ユーザの興味で切り出して提供することが可能である。しかし、かかるソフトウェアでは、各社別の定義ファイルをそれぞれ用意し、これを使用することで情報を取り出すことから、各社別の定義ファイルを十分に作り込まないと使用することができず、汎用性に欠けるものであった。

【０００９】

更に、現在、Ｗｅｂクリッピングサービス等で、ユーザの指定によって、任意のウェブサイトの位置から情報取得を可能とするものが存在する。ここでは、ページのＤＯＭ(Document Object Model)構造に着目し、XPathを利用して、指定された位置を、自動的に、指定された期間ごと若しくは変更があったときにクリッピングすることができる。しかしながら、ページ全体の構造や、レイアウトが

変わった場合に、ＤＯＭ構造も変化してしまい、自動的にクリッピングすることが困難となる。

【００１０】

本発明は、以上のような技術的課題を解決するためになされたものであって、その目的とするところは、例えばＷｅｂ上に公開されている様々な領域のカタログ等を、自動的に切り出すことにある。

また他の目的は、切り出された同一項目を合算して、例えば一つの表にしてユーザに提供することにある。

更に他の目的は、広範囲な領域での合算に対応することにある。

【００１１】

【課題を解決するための手段】

かかる目的のもと、本発明は、コンピュータがそのまま解釈できるように構造化されていない文書(データファイル)から、語彙と語彙との関係を定義したオントロジを利用して解析することで、Ｗｅｂ上にばらばらに存在する既存の各社カタログ等の中から、有用な情報を自動的に取り出し、同じ意味を持つ情報等を合算させ、例えば比較表等、ユーザに対して利便性のよい形で合算された情報を提供している。即ち、本発明が適用される情報収集システムは、構造化されていない複数のデータファイルをネットワークを介して取得する取得手段と、この取得手段により取得された複数のデータファイルに対し、所定の切り出しルールと用語の関係記述であるオントロジとを利用して解析する解析手段と、この解析手段による解析に基づいて、複数のデータファイルから必要な情報を抽出する抽出手段とを含んでいる。

【００１２】

ここで、この取得手段により取得されるデータファイルは、機械(コンピュータ)により読み取ってそのまま２次処理を行うことのできない、所謂構造化されてないテキスト、音、絵等を含む。特に、この取得手段は、ユーザの興味に基づくＵＲＬ(Uniform Resource Locators)を用いてＨＴＭＬ(Hypertext Markup Language)で書かれた文書を取得し、この解析手段は、特定のタグ情報を用いて文書を解析することを特徴とすることができる。この特定のタグ情報としては、ＨＴ

ＭＬのTableタグやListタグ等が挙げられる。尚、切り出しルールおよびオントロジは、ユーザ入力に従って適切なものを選択することができる。このとき、ユーザからの入力データにオントロジ特定データと切り出しルール特定データが含まれている場合の他、何らかのユーザの興味を示す入力データに基づいて、切り出しルールやオントロジを選択する場合もある。

【００１３】

また、この解析手段による解析に用いられる所定の切り出しルールは、カタログおよび/または仕様情報を構成する特徴をルール化したものとすることができる。更に、この解析手段は、用語の異なる複数のデータファイルに対し、オントロジを利用して横断的に内容を解析することを特徴とすることができ、また更に、抽出手段により抽出された情報を再構築し、この情報の中から同値関係をまとめてユーザ端末に提供する提供手段を含むことができる。

【００１４】

また、対象ごとに異なったオントロジを格納するオントロジ格納手段を備え、解析手段は、このオントロジ格納手段から所定のオントロジを読み出して解析することを特徴とすれば、プログラムに大きな変更を加えずとも、様々な分野の情報収集、解析に対応できる点から好ましい。

【００１５】

一方、本発明が適用されるアプリケーションサーバは、ユーザの興味に関する情報を受信するユーザ要求受信部と、このユーザ要求受信部より受信した情報に基づいて、複数のサイトからＨＴＭＬ文書を取得するＨＴＭＬ取得部と、ユーザ要求受信部より受信した情報に基づいてオントロジを読み出し、語彙情報を得る語彙情報処理機構と、ＨＴＭＬ取得部から取得したＨＴＭＬ文書に対し、語彙情報処理機構から提供される語彙情報に基づき、ＨＴＭＬ文書のタグを頼りに抽出データオブジェクトを取り出す抽出位置情報特定部とを含んでいる。

【００１６】

ここで、ＨＴＭＬ取得部から取得したＨＴＭＬ文書に対して切り出し処理を施すための切り出しルールを提供する切り出しルール処理機構、公理ルールに基づいて推論演算を実行する推論処理機構を更に備え、この抽出位置情報特定部は、

切り出しルール処理機構から提供される切り出しルールに基づいて、また、推論処理機構にて実行される推論演算に基づいて、抽出データオブジェクトを取り出すことを特徴とすることができる。

【００１７】

また、抽出位置情報特定部によって取り出された複数の抽出データオブジェクトに対して合算処理を施す情報整理集約部と、情報整理集約部による合算処理の結果に基づいてテーブルおよび/またはリストを生成する合算結果オブジェクト生成部と、この合算結果オブジェクト生成部により生成された合算結果オブジェクトを送信するユーザ要求送信部とを更に備えたことを特徴とすれば、ユーザに対して利便性のよい形で合算結果を提供できる点で優れている。

【００１８】

更に、本発明が適用される情報収集方法は、ネットワークに接続されたコンピュータにおいて、構造化されていない複数のデータファイル(ＨＴＭＬの文書)をネットワークを介して取得するステップと、ネットワークを介して取得されるＨＴＭＬの文書からTableタグまたはListタグに基づいて情報を抽出するステップと、取得され情報が抽出された複数のデータファイルに対し、所定の切り出しルールと用語の関係記述であるオントロジとを利用して解析するステップと、解析された複数のデータファイルから有用な情報を抽出するステップと、抽出された有用な情報をユーザにとって利便性のよい形で再構築するステップとを含んでいる。ここで、この解析するステップは、カタログおよび/または仕様情報を構成する特徴をルール化した切り出しルールを用いてテーブルの位置決めを行うステップと、位置決めされたテーブルの見出しがユーザにより指定されたカテゴリで一般に使用されている語彙であるか否かの語彙情報を定義するオントロジによって語彙のゆれを平準化するステップとを含むことを特徴とすることができる。

【００１９】

他の観点から捉えると、本発明が適用される情報収集方法は、インターネットに接続されたコンピュータにおいて、ユーザの興味に関する情報を受信するステップと、ユーザの興味に基づき、インターネットを介して複数の文書を取得するステップと、格納されている複数のオントロジから、ユーザの興味に基づいて特

定のオントロジを選定するステップと、選定された特定のオントロジを用いて、取得された複数の文書に対して横断的に内容を解析し、有用な情報を抽出するステップとを含んでいる。

【００２０】

更に、本発明が適用される情報収集方法は、ネットワークに接続されたコンピュータにおいて、関連性のある内容に対して各々異なった語彙にて表現された情報を含む複数のＷｅｂページを取得し、取得した複数のＷｅｂページからTableタグまたはListタグに基づいて情報を抜き出し、抜き出した情報に対して、語彙と語彙との関連付けを示すオントロジに基づき複数のＷｅｂページの異なる語彙に対して横断的に情報を解析し、解析された情報を合算し、合算結果をユーザ端末に対して送信することを特徴としている。ここで、この合算は、各Ｗｅｂページで異なる語彙に対して、上位下位概念の処理および/または関係処理を施して項目のマッチングを行うことを特徴とすることができる。

【００２１】

更に本発明は、ネットワークに接続されたサーバとして機能するコンピュータによって実行されるプログラムとして把握することができる。このプログラムは、構造化されていない複数のデータファイルをネットワークを介して取得する機能と、取得された複数のデータファイルに対し、所定の切り出しルールと、用語の関係記述であるオントロジと、所定の公理ルールに基づく推論演算とを利用して解析する機能と、解析された複数のデータファイルから有用な情報を抽出する機能と、抽出された有用な情報をユーザにとって利便性のよい形、例えば、関連性のある語彙と値について同値関係を処理し、更に新たな関係を挿入して情報を再構築する機能とをコンピュータに実現させている。

【００２２】

また、本発明が適用されるプログラムは、ユーザの興味に関する情報に基づき、インターネットを介して複数の文書を取得する機能と、格納されている複数のオントロジから、ユーザの興味に基づいて特定のオントロジを選定する機能と、選定された特定のオントロジを用いて、取得された複数の文書に対して横断的に内容を解析する機能とをコンピュータに実現させる。

【００２３】

更に、本発明が適用されるプログラムは、関連性のある内容に対して各々異なった語彙にて表現された情報を含む複数のＷｅｂページを取得する機能と、取得した複数のＷｅｂページからTableタグまたはListタグに基づいて情報を抜き出す機能と、抜き出した情報に対して、語彙と語彙との関連付けを示すオントロジに基づき複数のＷｅｂページの異なる語彙に対して横断的に情報を解析する機能と、解析された情報を合算する機能とをコンピュータに実現させる。

【００２４】

これらのプログラムとしては、コンピュータを顧客に対して提供する際に、例えばサーバ等の装置の中にインストールされた状態にて提供される場合の他、コンピュータに実行させるプログラムをコンピュータが読取可能に記憶した記憶媒体にて提供する形態が考えられる。この記憶媒体としては、例えばフロッピーディスクやＣＤ－ＲＯＭ媒体等が該当し、フロッピーディスクドライブやＣＤ－ＲＯＭ読取装置等によってプログラムが読み取られ、フラッシュＲＯＭ等にこのプログラムが格納されて実行される。また、これらのプログラムは、例えば、プログラム伝送装置によってネットワークを介して提供される形態がある。このプログラム伝送装置としては、例えば、ホスト側のサーバに設けられ、プログラムを格納するメモリと、ネットワークを介してプログラムを提供するプログラム伝送手段とを備えている。

【００２５】

【発明の実施の形態】

以下、添付図面に示す実施の形態に基づいて本発明を詳細に説明する。

図１は、本実施の形態が適用される情報収集システムの全体構成を示した図である。図１に示す情報収集システムは、例えばＰＤＡ(Personal Digital Assistant)やノートＰＣなどのネットワーク接続が可能なユーザ端末１１、各社ごとに設けられ各種カタログや情報からなるＷｅｂページを提供するＷｅｂサーバ１２、本実施の形態における情報収集サービスをユーザ端末１１に提供するＷｅｂアプリケーションサーバ２０を備え、これらがインターネット１０を介して接続されている。尚、Ｗｅｂアプリケーションサーバ２０だけを捉えて、狭義の情報収

集システムとして把握することも可能である。かかる場合等において、「システム」の文言は、各機能が筐体を同じくするか、所定のネットワークを介して接続されているかを問うものではない。

【００２６】

Ｗｅｂアプリケーションサーバ２０は、ユーザ端末１１からユーザの興味の登録を受け、情報収集サービスに対する最初のアクセスページを提供するポータルサーバ２１、各社のＷｅｂサーバ１２からインターネット１０を介して情報の収集を行うインフォメーション/サービス・モニタ・エージェント２２、語彙と語彙との関連付けであるオントロジ(Ontology)をデータベースに格納し、語彙情報群を提供するオントロジサーバ２３、ポータルサーバ２１を介して得られたユーザ要求から情報収集処理を実行しユーザ端末１１に提供する情報配信システム２４を備えている。この情報配信システム２４は、ユーザ端末１１からユーザが登録した自身の興味と収集した情報とが合致しているか否かを調べている。オントロジサーバ２３は、対象ごとに異なったオントロジ(例えば、ノートＰＣオントロジ、デジタルカメラオントロジ、不動産オントロジ等)をそれぞれのデータベースに格納しており、対象ごとにオントロジを入れ替えるように機能している。また、情報配信システム２４では、例えば「Ａ社の株価が１００を超えるという情報があれば通知して欲しい。」といったユーザの興味を登録すると、インフォメーション/サービス・モニタ・エージェント２２によって収集された情報を調べ、興味に一致している情報があれば合致しているという結果を返している。

【００２７】

ここで、理解を容易にするために、本実施の形態における情報収集処理の概要について説明する。一般に、インターネット１０を介して得られるＨＴＭＬで記述された情報は、ユーザ端末１１のユーザ等に対して視覚的に表現するための効果を狙って記述されており、コンピュータに対しては非構造化(構造化されていない)のデータファイルであると言える。そのために、インターネット１０上の複数の情報を比較(収集・整理)するには、多くの手間が必要となる。即ち、データ構造を簡単に扱える形式をもっておらず、ＨＴＭＬで書かれたこれらの情報では、各ページの中でどこに情報があるのかを機械的に把握することが難しく、機

械的に２次処理を施して情報を取り出すことは難しい。また、同じ概念が異なる語彙で表現されている場合も多く、ユーザにとって有用な情報を機械的に抽出することが困難である。本実施の形態では、Ｗｅｂ上に公開されている様々な領域のカタログ等を電子的に配布し、配布されたパンフレット・カタログを自動的に切り出し、同一項目を合算して一つの表にすることで、ユーザによる比較を容易にしている。また、本実施の形態では、各領域ごとのカタログ等に記載されている表に対し、切り出しルール、語彙、概念体系(オントロジ)を挿げ替えることで、広範な領域での合算に対応することができる。

【００２８】

図６および図７は、Ｗｅｂ上に公開されているカタログの一例を示した図である。ここでは、ＰＣを販売しているメーカの各Ｗｅｂサーバ１２から提供されるＷｅｂページの例を示している。図６に示すカタログでは、コンピュータの入出力や命令の実行などを行うＣＰＵを「プロセッサ」と呼び、各機種ごとに、この「プロセッサ」の仕様が表現されている。一方、図７に示すカタログでは、この部分を「ＣＰＵ」と呼び、各機種ごとにその仕様が表示されている。従来では、これらのホームページ(ＨＰ)から得られたカタログについて、購入等の際に、ユーザが自ら目で見て手作業で比較することが必要であった。

【００２９】

図８は、本実施の形態における合算表示例を示した図である。ここでは、図６のＨＰに示す商品情報と図７に示すＨＰの商品情報とをまとめ、例えば、図６に示す「プロセッサ」と図７に示す「ＣＰＵ」とを「プロセッサ」の項目で合算して、表示している。具体的には、語彙と語彙との関連付けである「オントロジ」を用い、今まで意味付けの概念が存在しなかったテーブルを、Ｗｅｂ上から、以下実施例で述べる手法を適用することで切り出しを行う。そして、各テーブルの欄にオントロジを用いて、上位、下位概念の関係、類義語、反意語、および論理演算、述語関係による推論を施し、意味を類推することにより、各社ごとの表を一つの表に合算している。即ち、オントロジを用いて各テーブルに意味を持たせ、その意味に応じてそのテーブルを切り出し、同じ意味を有するもの同士を合算している。これにより、各社ごとのその機能を表す単語が異なっていても、意味

付けによって自動的に同じ物であると判断し、例えばＣＰＵとプロセッサとは同じ物として合算している。この合算された表を参照することによって、ユーザは、各社まちまちの単語を用いて表現されていた情報に対して、例えば統一した用語を用いて容易に比較することができる。

【００３０】

このとき、本実施の形態では、各サイトごとに作り込みを行ってはいない。そのために、例えば、ノートＰＣ用オントロジ、デジタルカメラ用オントロジ、不動産オントロジ等、各対象領域ごとにオントロジを定義でき、プラグインすることで動的に対処できる。この技術によれば、各テーブルの値にオントロジ操作を施すことで、例えばＨＴＭＬ(Hypertext Markup Language)でのテーブルなどのように、「人間には表の意味が理解できるが機械にはただの表示の手段に過ぎず、表の各欄の意味は理解できない。」という言語から、機械にも理解できるようなＸＭＬ(Extensible Markup Language)やＲＤＦ(Resource Description Framework)といった形式に自動的に変換することが可能となる。また、具体的な応用例としては、このように各ＨＴＭＬのテーブルに意味を付けることができると、例えば、プログラム製品のＩＮＳ(Intelligent Notification Services)を使うことにより、予め登録しておいたユーザの興味のある事象と、既存のＷｅｂページの内容とが一致しているか、といった定量的な比較が可能となり、ユーザの興味が一致した場合に、ユーザに通知するように構成することも可能である。

【００３１】

ここで、「オントロジ」とは、意味情報を表現するための方法の一つであり、概念同士の関係やそれらを解釈するための論理的なルールを定義する文章の集合である。例えば、「日曜日午前大和内科」という内容を検索するとする。現状のＷｅｂでは、ＨＴＭＬからそのままの単語を取り出して検索結果としており、多くの検索ゴミが発生していた。一方、「オントロジ」では、ａ．大和は市の名前であること、ｂ．病院には内科、外科、耳鼻科があること、ｃ．病院には診察日や診察時間があること、等のそれらを解釈するための論理的なルールが定義され、この文章の集合から検索結果を得ることができる。その結果、検索ゴミを少なくすることが可能となる。本実施の形態では、抜き出したテーブルにこの「オン

トロジ」を用い、各ページで異なる語彙に上位下位概念、関係処理を施し、項目のマッチングを行い、言葉のゆれなどの形態素にまつわる処理を行っている。このとき、色々な領域(例えば保険、株式、病院、不動産、車、ＰＣ等)に対する「オントロジ」を用意することで、色々な領域に対して応用することができる。

【００３２】

次に、かかる情報収集方法を実現するための構成について、以下に詳述する。

図２は、本実施の形態が適用される情報配信システム２４の機能構成を示したブロック図であり、図１に示すＷｅｂアプリケーションサーバ２０にて実行される。ここでは、ユーザの興味に関する情報を受信するユーザ要求受信部３１、ユーザ要求受信部３１により指定されたＵＲＬからＨＴＭＬの文書を取得するＨＴＭＬ取得部３２、ＨＴＭＬのテーブルに着目して、抽出するデータが含まれるテーブル(位置)を特定する抽出位置情報特定部３３、得られた複数のサイトからの情報を合算する情報整理集約部３４、合算した情報(合算処理オブジェクト)をテーブル等の指定された表示形式(合算結果オブジェクト)に変換してこれらを表示する合算結果オブジェクト生成部３５、合算結果をユーザに提供するユーザ要求送信部３６を備える。また、ユーザ関心表現式により関連する切り出し(抽出)ルール群をロードする切り出しルール処理機構４１、ユーザ関心表現式により関連するオントロジをロードする語彙情報処理機構４２、抽出位置情報特定部３３や情報整理集約部３４から呼ばれて様々な推論演算を実行する推論処理機構４３を有している。

【００３３】

まず、ユーザ要求受信部３１では、ユーザの興味を適切に表現するコンポーネントとして、例えばＳＱＬ(Structured Query Language)等で書かれたユーザ関心表現式を受信する。このユーザ関心表現式は、ノートＰＣの例では「価格が１５万円以下のノートブックを表示」といった具合になる。また、他の方法として、例えばユーザによる特定のキーワード入力を受け、所定のプログラムがこのキーワードからＵＲＬ(Uniform Resource Locators)とオントロジ種別とを特定し、ユーザ関心表現式として扱うことができる。即ち、テキスト入力を受けた後、全文検索エンジンから見込みのある対象ＵＲＬを得ることで、ユーザ関心表現式

を作成する。例えば、ユーザからの指定や検索によって、以下のようなＵＲＬ、オントロジ種別を得ることができる。

```
UserInterest
  URL1 http://xxx.yyy
  URL2  HYPERLINK "http://yyy.xxx" http://yyy.xxx
  TargetSpec NotePCSpecOntology
                (DegitalCameraSpecOntlogy,RealEstateSpecOntology)
```

【００３４】

ＨＴＭＬ取得部３２は、ユーザ要求受信部３１から上述したようなＵＲＬを取得する指定ＵＲＬ取得部５１、この得られたＵＲＬからＨＴＭＬ部分を解析するＨＴＭＬ解析部５２を備えている。得られた情報位置式ＵＲＬとしては、例えば、http://www.somecompany.com/products/notepc/newproduct.html 等である。まず、最初にＷｅｂアプリケーションサーバ２０側にて取得された状態としては、ＨＴＭＬオブジェクト(ＨＴＭＬの構文解析木(ツリー構造))として１ページを丸ごと取得した状態にある。また、ＤＯＭ(Document Object Model)によって、ＨＴＭＬのデータ構造解析を行い、タグ情報が取得される。ＨＴＭＬ解析部５２では、例えばＡＰＩ(Application Program Interface)を利用して、ＨＴＭＬオブジェクトからテーブル部分だけの情報、即ち、Tableオブジェクト(ＨＴＭＬ構文解析木のサブセット)を抜き出している。尚、リストについても同様に、Listタグの中のツリー構造に対して同様な手法を用いて抜き出すことが可能である。

【００３５】

抽出位置情報特定部３３では、切り出しルール処理機構４１、語彙情報処理機構４２、推論処理機構４３が呼び出され、抽出データオブジェクトが取り出される。そのために、この抽出位置情報特定部３３は、ＨＴＭＬ取得部３２によって得られたＨＴＭＬオブジェクトから、<UL>、<OL>、<LI>等のリスト構造から抽出データオブジェクトを取り出すリスト構造抽出部５３、テーブル構造から抽出データオブジェクトを取り出すテーブル構造抽出部５４、Tableタグが入れ子である場合に、更に内部のTableタグで囲まれた部分を抽出する情報提示位置特定部５５を有している。つまり、テーブル構造やリスト構造を構文解析するＨＴＭＬ

解析部５２でオブジェクトを抽出可能な状態にし、リスト構造抽出部５３またはテーブル構造抽出部５４で、意味のあるオブジェクトを特定し、抽出データオブジェクトとして取り出している。抽出される部分の例としては、例えば、図７に示すカタログのページからは、

Table1

Vocabulary:CPU　Value: Mobile CPU III

Vocaburary:HDD　Value: 20GB

：　　：

例えば、図６に示すカタログのページからは、

Table2

Vocabulary:プロセッサ　Value: PPP PC

Vocaburary:ハードディスクドライブ　Value: 15GB

：　　：

：

といったようなTableタグで囲まれた部分が抽出され、これらがフラットに並べられる。尚、Formタグを手がかりに解析する際は、ユーザからの興味やキーワード入力等のプリファレンスをもとに、Formエレメントのインプットタグに自動的にフォームの入力データを挿入し、リクエストを自動サブミットし、その結果、レスポンスとして得られるＨＴＭＬよりTableタグ、Listタグを抽出し、情報を収集している。

【００３６】

切り出しルール処理機構４１は、ルール群を管理するルール群管理機構６３、該当するルール群を所定のメモリからロードする切り出しルールロード部６４を備え、多くの切り出し（抽出)ルールからなる切り出しルール群が準備される。この切り出しルール群としては、例えば、

・一行目には全て同じ項目となる場合が多い。

・仕様に関係する語彙は一桁目に来る。

・一桁目(項目桁)とそれに対応する右側にあるカラムとはある関係を持つ。

・空白のセルはある程度より多くない。

・ＣＰＵのカラムに対応するカラムには重量を表す１ｋｇはこない。

等の複数のルールが存在し、抽出位置情報特定部３３は、これらのルール群を参照して、抽出位置情報を特定している。

【００３７】

語彙情報処理機構４２は、語彙情報群を管理する語彙情報管理機構６５、所定のメモリから語彙情報をロードする語彙情報ロード部６６を備え、ユーザ関心表現式よりオントロジをロードして(例えば、図１に示すオントロジサーバ２３から所望の(対象の)オントロジを読み出して)語彙情報群を得ている。語彙情報の例として、例えば、各社のＰＣを比較する際に使われるオントロジでは、以下のようなものがある。

Class CPU sameAs プロセッサ

Class プロセッサ sameAs CPU

Class キャッシュメモリ

Class L2キャッシュ subClassOf キャッシュメモリ

Class 重量　subClassOf

　unionOf 本体質量

　　バッテリ質量

ここで、“sameAs”は、「～と同じ意味」、“subClassOf”は、「上位下位の関係」、“unionOf”は、「含む」である。例えば、オントロジを用いて「重量」は「本体重量」＋「バッテリ」という関係を定義することで、ユーザに有用な情報に変換することが可能となる。

【００３８】

このように、語彙情報処理機構４２によって提供される語彙情報は、語彙間の関係を持ち、例えば、上位、下位概念関係、同義、反義、類義といった一般的な関係から、その語彙特有の関係(物理的関係、時系列的関係、単位系)、および語彙情報定義者の個別定義による種々の関係定義などを持つことができる。また、そのような語彙情報は、基本概念を構成するものと、領域に応じて作成するものがあり、領域に応じて作成されるものは、基本概念を構成するものをベースにし、他の領域の語彙情報を参照することもできる。

【００３９】

推論処理機構４３は、推論演算を実行する推論エンジン６８、推論エンジン６８の実行を制御する推論エンジン実行制御機構６７、所定のメモリから公理ルール群をロードする基本(公理)ルールロード部６９を備え、推論エンジン６８の受け付けるルール記述形式により記述されたルール群である公理ルールを用いて推論処理を実行している。ここでは、オントロジをセマンティック(Semantic)実行するために推論エンジン６８を使用し、駆動ルールが実装されている。例えば、事実のみから三段論法が実行され、Ｗｅｂ上に散在する事実(オントロジ言語により記述されたメタ情報)から推論するために、定言三段論法が実装される。この定言三段論法としては、例えば、

(大前提)　全ての人間は死すべきものである。

(小前提)　ソクラテスは人間である。

→(結論)　ゆえにソクラテスは死すべきものである。

といったものが挙げられる。

【００４０】

論理型言語による通常の三段論法は、事実(定言)と、含意、もし～ならば(仮言)からなる混合仮言三段論法で表現される。論理型言語での例では、

```
mortal(X) : - man(X)     (大前提)  仮言(条件)
man(socrates).           (小前提)  定言(事実)
? - mortal(socrates).  →  yes.  (結論)
```

【００４１】

定言三段論法の実装として、推移律の実装では、以下のようになる。

```
/*TransitiveProperty*/
if pv( type,?p, TransitiveProperty) and
   pv(?p,?x,?y) and
   pv(?p,?y,?z)
then
   pv(?p,?x,?z)
```

このようにして、矛盾したものを外し、同値のものを得る等、推論処理機構４３

では、事実から新しい事実を三段論法によって導出するための公理ルールを提供している。

【００４２】

このように、推論処理機構４３では、語彙情報処理機構４２によって上記の様に定義された語彙情報における関係を操作するために、推論エンジン６８を使用し、様々な関係における論理演算をルールとして実装している。例えば、矛盾した語彙の発見、包含関係の発見、三段論法による新事実の発見等により、カタログ、仕様情報を構成するテーブルなどの切り出しの精度を上げ、且つ、複数のページから切り出された情報を付き合せる際にも同様の手法を適用し、情報の整理、集約の自動実行を可能としている。尚、定言三段論法以外に、反対(inverse)関係や矛盾(disjoint)関係等を駆動するための公理ルールがある。本実施の形態では、基本的な公理ルールによって、オントロジで定義された関係を他の関係との間に適応して新事実、矛盾等が推論できるように、推論処理を駆動している。

【００４３】

情報整理集約部３４では、合算処理を行う情報合算部５６、合算対象の位置決めの特定を行う合算対象位置決め特定部５７を有し、抽出位置情報特定部３３で取り出された抽出データオブジェクトから合算処理オブジェクトを生成している。この合算処理を行う際、語彙情報処理機構４２および推論処理機構４３が呼び出され、オントロジがそれぞれの語彙について対応付けられ、推論を用いてその結果が集約できるように構成されている。この合算処理オブジェクトは、語彙と値との対応付けを横断的に行い、同値関係を処理し、更に新たな関係も挿入されたものである。その例としては、

```
Object
  Entry 1
    Class:CPU OriginalVoc: CPU Value: Mobile CPU III
    Class:HDD OriginalVoc: HDD Value: 20GB
          :                    :              :
  Entry 2
    Class:CPU OriginalVoc: プロセッサ  Value: PPP PC
```

Class:HDD OriginalVoc: ハードディスクドライブ　Value: 15GB

：　　：　　：

：

のようなデータ構造である。ここでは、「ＣＰＵのオリジナルボキャブラリとしてはＣＰＵがある。」や、「ＣＰＵにてオリジナルボキャブラリではプロセッサとなっていた。」といったようなオブジェクトが生成される。

【００４４】

このようにして、情報整理集約部３４では、得られた２つのサイトからの例えばノートＰＣの情報が合算される。例えば、Ａ社ＰＣのＣＰＵがｘｘｘ、Ｂ社ＰＣのプロセッサがｙｙｙというデータが、ここで、Ａ社ＰＣのプロセッサ(つまりＣＰＵ)がｘｘｘ、Ｂ社ＰＣのプロセッサ(つまりＣＰＵ）がｙｙｙ、というデータとして、互いに比較対象として並べることのできる位置に再配置され、合算処理オブジェクトとして保持される。

【００４５】

合算結果オブジェクト生成部３５では、合算結果テーブル生成部５８、合算結果リスト生成部５９を備え、情報整理集約部３４から得た合算処理オブジェクトに対して、ユーザに対して見やすい形で合算結果を提供できるようにテーブルおよび/またはリストを生成し、合算結果オブジェクトを生成している。

【００４６】

ユーザ要求送信部３６では、合算結果オブジェクト生成部３５により生成された合算結果オブジェクトから合算結果ＨＴＭＬを生成する合算結果ＨＴＭＬ生成部６１、生成されたＨＴＭＬをユーザ要求受信部３１にて要求を受信したユーザに対して送信するユーザ要求結果送信部６０を備え、図８に示すような比較表がユーザ端末１１のユーザに提供される。

【００４７】

次に、フローチャートを用いて、これらの処理について説明する。

図３は、図２のブロック図に示す各機能によって実行される全体処理の流れを示したフローチャートであり、上位の概念からの処理を説明している。まず、ＨＴＭＬ取得部３２は、ユーザ要求受信部３１からの情報位置式に指定されたＵＲ

Ｌへアクセスし(ステップ１０１)、抽出位置情報特定部３３は、ＨＴＭＬ取得部３２により取得された比較対象のＨＴＭＬからテーブルを全て取得する(ステップ１０２)。切り出しルール処理機構４１では、対象物に対する切り出しルールがロードされる(ステップ１０３)。語彙情報処理機構４２では、対象物に対するオントロジがロードされ、テーブルの切り出しに使用される(ステップ１０４)。抽出位置情報特定部３３では、切り出しルール処理機構４１でロードされた切り出しルールや語彙情報処理機構４２にてロードされたオントロジ、推論処理機構４３によってロードされる公理ルール等を用いて、これらの取得したテーブルより対象物の仕様のテーブルの抜き出しが行われる(ステップ１０５)。ここで次の比較対象がまだあるか否かが判断され(ステップ１０６)、次の比較対象がある場合には、ステップ１０１へ戻り、次の比較対象がない場合には、語彙情報処理機構４２にて対象物に対するオントロジがロードされ、ステップ１０９におけるテーブルの合算に使用される(ステップ１０７)。また、推論処理機構４３では、推論エンジン６８により、現在の関係を用いて新たな関係が作成される(ステップ１０８)。そして、語彙情報処理機構４２にてロードされたオントロジおよび推論処理機構４３により作成された新たな関係等を用いて、情報整理集約部３４にて同一項目の合算処理が行われ、合算結果オブジェクト生成部３５にて合算結果のオブジェクトが生成される(ステップ１０９)。その後、ユーザ要求送信部３６によって合算結果がユーザに表示され(ステップ１１０)、全体の処理が終了する。

【００４８】

次に、実施の形態にて説明した例を用いて、処理の流れを説明する。

図４は、本実施の形態が適用される処理を更に詳述したフローチャートである。まず、ユーザ要求受信部３１では、ユーザの要求(興味)が受信される(ステップ２０１)。この受信されたユーザの要求に基づいて、ＨＴＭＬ取得部３２では、ユーザの興味のあるＵＲＬにアクセスし、ＨＴＭＬが取得される(ステップ２０２)。このとき、例えばテーブルのあるＵＲＬは、予め指定されているものとすることができる。抽出位置情報特定部３３では、得られたＨＴＭＬがＤＯＭで解析され(ステップ２０３)、テーブルタグの部分のみが切り出される(ステップ

２０４)。ここでテーブルタグが入れ子か否かが判断され(ステップ２０５)、入れ子である場合には、更に内部のテーブルタグで囲まれた部分を抽出し(ステップ２０６)、入れ子が残っている間はステップ２０５とステップ２０６が繰り返される。

【００４９】

ステップ２０５にてテーブルタグが入れ子ではない場合には、例えばノートＰＣ仕様の切り出しルール、オントロジが、切り出しルール処理機構４１および語彙情報処理機構４２によりロードされているか否かが判断される(ステップ２０７)。作成されていない場合には、切り出しルール処理機構４１にて、前述したような切り出しルールが選択され、ロードされて、例えばノートＰＣ仕様部分のテーブルが切り出される(ステップ２０８)。また、語彙情報処理機構４２では、語彙情報(必要なオントロジ、例えばノートＰＣオントロジ等)が選択され、ロードされる(ステップ２０９)。また、推論処理機構４３では、推論エンジン６８が使用され、駆動ルールが実装されて、語彙の関係付けが行われて(ステップ２１０)、ステップ２０７の判断に戻る。ここでは、例えば、“unionOf”がきたらその合計を計算する等、事実のみから三段論法等が実行される。このように、オントロジが選択され、選択されたオントロジが用いられることで、例えば、各テーブルの欄にオントロジを用いて、上位、下位概念の関係、類義語、反意語、および論理演算、述語関係による推論を施し、意味を類推することにより、各社ごとの表を一つの表に合算することができる。推論エンジン６８をノートＰＣに適用した場合には、例えば、「重量」は「本体重量」＋「バッテリ」であるという事実(オントロジ)について、実際に推論エンジン６８を使用して駆動される。例えば、「本体という用語とバッテリという用語があり、そのフィールドに重さを表す情報があれば、その２つを足して、重量というラベル付けをした事実とする。」という処理が実行される。

【００５０】

ステップ２０７にてノートＰＣ仕様のテーブルが作成されている場合には、抽出位置情報特定部３３では、オントロジ、切り出しルールを用いて、ノートＰＣ仕様のテーブルの切り出しが行われる(ステップ２１１)。内部的には、それらを

ベースにした評価関数(ルールがどの程度、真になっているか等)をもとに判断が行われる。この切り出しの後、情報整理集約部３４にて、各ノートＰＣ仕様のテーブルが比較できる状態に作成されているか否かが判断される(ステップ２１２)。例えば、同じ項目同士があるかどうか、同じ項目らしきものについて違う言葉で書かれているか否か等について、判断できる状態にテーブルが作成されているか否かが判断される。作成されていない場合には、語彙情報処理機構４２にアクセスし、オントロジを語彙に用い(ステップ２１３)、また、推論処理機構４３にアクセスし、推論エンジン６８を使用して、語彙の同値関係の付与等、新たな関係が作成され(ステップ２１４)、ステップ２１２の判断に戻る。ステップ２１２にてテーブルが比較できる状態に作成されている場合には、情報整理集約部３４にて、各ノートＰＣ仕様が項目毎に合算され、合算結果オブジェクト生成部３５にて合算結果のテーブルが生成される(ステップ２１５)。その後、ユーザ要求送信部３６にて、出来上がった合算結果がＨＴＭＬでテーブル形式に直され、ユーザ端末１１に表示され(ステップ２１６)、処理が終了する。尚、ステップ２１１のテーブルの切り出しに際して、比較できないテーブルとしては、例えばノートＰＣにおける合算の場合の標準的な用語に各フィールド項目が正規化されていない状態にあるものが該当する。標準的な用語は、語彙情報群により用途ごと(この例ではノートＰＣごと)に予め決定されている。例えば、ＣＰＵという用語が語彙情報群の標準ノートＰＣのスペックとして定義されている場合、ステップ２１３およびステップ２１４の処理によって、プロセッサという用語が使用されているフィールド名がＣＰＵというフィールド名に変換される。

【００５１】

図５は、ユーザ端末１１に対する表示を更に詳述したフローチャートである。ユーザ要求受信部３１にて、テーブルを有するＵＲＬが予め指定されている場合に、ＨＴＭＬ取得部３２では、比較対象のＨＴＭＬからテーブルが全て取得される(ステップ３０１)。次に、抽出位置情報特定部３３では、取得したテーブルよりノートＰＣ仕様のテーブルの抜き出しが行われ(ステップ３０２)、次の比較対象があるかどうかが判断される(ステップ３０３)。次の比較対象がある場合には、ステップ３０１に戻り、次の比較対象がない場合には、情報整理集約部３４に

てノートＰＣ仕様のテーブルが合算される(ステップ３０４)。

【００５２】

その後、ユーザ関心表現式から、ユーザの興味のあるもののみが抽出されたか否かが判断され(ステップ３０５)、そうではない場合には、情報整理集約部３４にて、内容を全て合算してユーザに表示し(ステップ３０６)、処理が終了する。このステップ３０５の「ユーザの興味あるもののみ抽出する」場合とは、例えばユーザ関心表現式でユーザが「ＨＤＤが１０Ｇバイト以上のノートＰＣの情報が欲しい。」と登録した場合、情報源から各ノートＰＣの情報が得られた後、情報の中からユーザの興味に合致したもののみを取り出すプロセスである。ユーザの興味あるもののみの抽出ではない場合には、得られた情報全てがユーザに届けられる。ステップ３０５でイエスの場合には、合算した結果が個々のＸＭＬファイルに分けられる(ステップ３０７)。そして、ユーザの興味と合致しているものがあるかどうかが判断され(ステップ３０８)、合致しているものがない場合にはそのまま処理が終了し、合致しているものがある場合には、合算結果オブジェクト生成部３５にて内容が合算されてユーザに表示され(ステップ３０９)、処理が終了する。

【００５３】

以上のように、カタログ、仕様情報は、テーブル、リスト形式で提示されている場合が多いが、従来技術では、ＨＴＭＬのTableタグ、Listタグでは表示形式を指定するだけであり、テーブル、リスト形式で提示された情報を収集、整理するためには、ブラウザに提示された情報を手動で集め、整理するしかなかった。また、表形式で示される情報の見出し(列、行に含まれる情報の見出し)が、情報提供者(ページ)によって異なり、単純に、機械的に整理することは難しかった。特に、テーブルタグはレイアウト情報としてページに多用されており、単純にテーブルタグから必要とする情報を抽出することは困難であった。本実施の形態では、どこに情報があるかを特定する機能を備え、指定されたページを読み込み、ページの情報が属するカテゴリ情報に対してユーザの指定を可能としている。また、ページに最適化された情報抽出ルールを使用し、情報が存在するテーブル、リストの位置決めを可能としている。尚、この情報抽出ルールでは、テーブル若

しくはリストといったＨＴＭＬ、ＴＡＧによる位置決めと、各カテゴリのページで使用されている語彙情報とを用いて、情報の位置決めが行われている。

【００５４】

また、テーブルの位置決めにおいては、レイアウト情報ではない、カタログ、仕様情報を構成する特徴をルール化し、位置決めの第一ステップとしている。また、この第一ステップにて、情報抽出を行ったテーブルにおいて、列見出し、行見出しが、ユーザによって指定されたカテゴリとして一般的に使用されている語彙であるかどうかを判断し、一般的な語彙情報をパターンとして定義し、ページ毎に異なる語彙の「ゆれ」について、語彙情報を使用して平準化し、テーブル特定の精度を上げている。尚、ページのカテゴリによる、Tableタグ、Listタグのレイアウト情報における使用パターンに応じ、このテーブルの位置決めを交換可能とし、また、カテゴリに応じた列見出し、行見出しに対して語彙情報を交換することで、多様なカテゴリに対応できる汎用的な機構を実現することもできる。このように本実施の形態では、あるページから必要な情報を複数抽出し、その複数の情報間の関係を利用し、情報の整理を行うことが可能である。

【００５５】

このように、本実施の形態では、構造化されていないデータファイルからオントロジを利用して解析し、有用な情報を抽出している。特に、インターネットで標準的に用いられているＨＴＭＬ言語で書かれた文書の解析を、Form、Tableタグなどをヒントに解析し、情報抽出を行っている。また、オントロジ(用語の関係記述)を使って、用語の異なる複数の文書に亘っても、横断的に内容を解析し、有用な情報の抽出を可能としている。また、解析した結果を用いて、利用者に更に利便性の良い形で情報を再構築して提示することもできる。特に、カタログ形式の情報の合算に応用できることや、オントロジを交換することでプログラムに大きな変更を加えることなく様々な種類のデータファイルに対応可能である。また、ＨＴＭＬからＸＭＬのような機械処理できる言語に変換することも可能である。

【００５６】

また、情報の抽出に際して、各Ｗｅｂページを作り込むわけではないことから

、例えば、ノートＰＣオントロジ、デジタルカメラ用オントロジ、不動産オントロジ等、抽出の対象毎にオントロジを入れ替えることで、動的にロードすることが可能となる。更に、各対象領域毎に抽出ルールをプラグインすることができ、色々な領域に対してプラグインを変えることで、適応することが可能となる。即ち、コアとなる部分は全て共通であることから、各Ｗｅｂページに対して作り直す必要がなく、保守性や生産性を向上させることができる。

【００５７】

さらに平均値や合計値なども算出することができる。また、ＨＴＭＬなどのメタ情報を持たない言語からＸＭＬなどのメタ情報を付加した言語への自動変換も可能である。本実施の形態における適用分野としては、Ｗｅｂサイトに関するＳＩ、ナレッジマネジメント、ポータルサイトへの付加価値なども適用分野として挙げられる。更に、意味把握機能を備えた知識表現におけるＷＷＷであるセマンティックウェブ(SemanticWeb)との相乗効果も期待できる。

【００５８】

以上、本実施の形態によれば、異なる用語を含んでいる複数の文書に対して、横断的に内容を解析することが可能となり、同じ意味を持つ情報を抽出することができる。同様に、構造化されていない文書からも目的とする情報を得ることが可能となる。また、解析した結果を合算し、比較表を作成することによって、ユーザにさらに利便性のよい形で情報を提供することができる。更に、オントロジを差し替えることで、プログラムに大きな変更を加えることなく、様々な分野に対応することが可能となる。

【００５９】

この本実施の形態における応用として、例えば、展示会等にて携帯情報端末等にパンフレット等を電子的に配布し、配布されたパンフレット・カタログに対して自動的に同類項目を合算するものが挙げられる。この合算した情報を比較が容易な表現形式等に変換したり、分類したりする機能を更に備え、変換結果や分類結果を携帯情報端末上に表示したり、印刷できるようにすれば、ユーザが展示会等に行った際に多量のパンフレット等を持ち歩く代わりに、携帯情報端末等を利用して、容易に比較検討することができる。即ち、展示会等にて、ＸＭＬ等によ

って構造化され、ＲＤＦによりメタデータが付与された電子パンフレットやカタログを、ローカル若しくはリモート上にあるオントロジ情報に基づいて、同一項目を抽出し、表形式のレポートをユーザに提供することが可能となる。

【００６０】

更に、他の応用として、Ｗｅｂ上に多々ある不動産情報等について、今まではユーザが一つ一つのサイトを手作業で比較していたものを、本実施の形態の技術を用いてＷｅｂ上の表を切り出し、オントロジ操作を施し、ユーザの目的の物件を複数の不動産情報サイトから合算して表示させることも可能である。また、車の情報に関しても、現状のＷｅｂ上の情報では各社まちまちで、比較するには各社で独自に他社の情報をデータベースに持って比較することが必要であったが、同様な方法を用いることで、現在あるＷｅｂページを使ってユーザには比較結果を届けることが可能となる。また、ショッピングやチケット、オークションといった、現在Ｗｅｂ上に存在するが、各社ごとに対応がまちまちで比較合算できない分野に有効である。更に、上述した実施の形態では、ＨＴＭＬのテーブルに注目したが、これをフォームに置き換えても利用可能である。このように、本実施の形態では、アドホックで未成熟なエリアに対してオントロジを適用し、汎用性のある方法を提供することで、アプリケーション開発の労力削減、オントロジ、ルールのモジュール化、プラグイン化による迅速な適用が可能となり、変更に強い情報検索システムを提供することが可能となる。

【００６１】

【発明の効果】

以上説明したように、本発明によれば、例えばＷｅｂ上に公開されている様々な領域のカタログ等を、自動的に切り出すことが可能となる。

【図面の簡単な説明】

【図１】　本実施の形態が適用される情報収集システムの全体構成を示した図である。

【図２】　本実施の形態が適用される情報配信システムの機能構成を示したブロック図である。

【図３】　図２のブロック図に示す各機能によって実行される全体処理の流

れを示したフローチャートである。

【図４】　本実施の形態が適用される処理を更に詳述したフローチャートである。

【図５】　ユーザ端末に対する表示を更に詳述したフローチャートである。

【図６】　Ｗｅｂ上に公開されているカタログの一例を示した図である。

【図７】　Ｗｅｂ上に公開されているカタログの一例を示した図である。

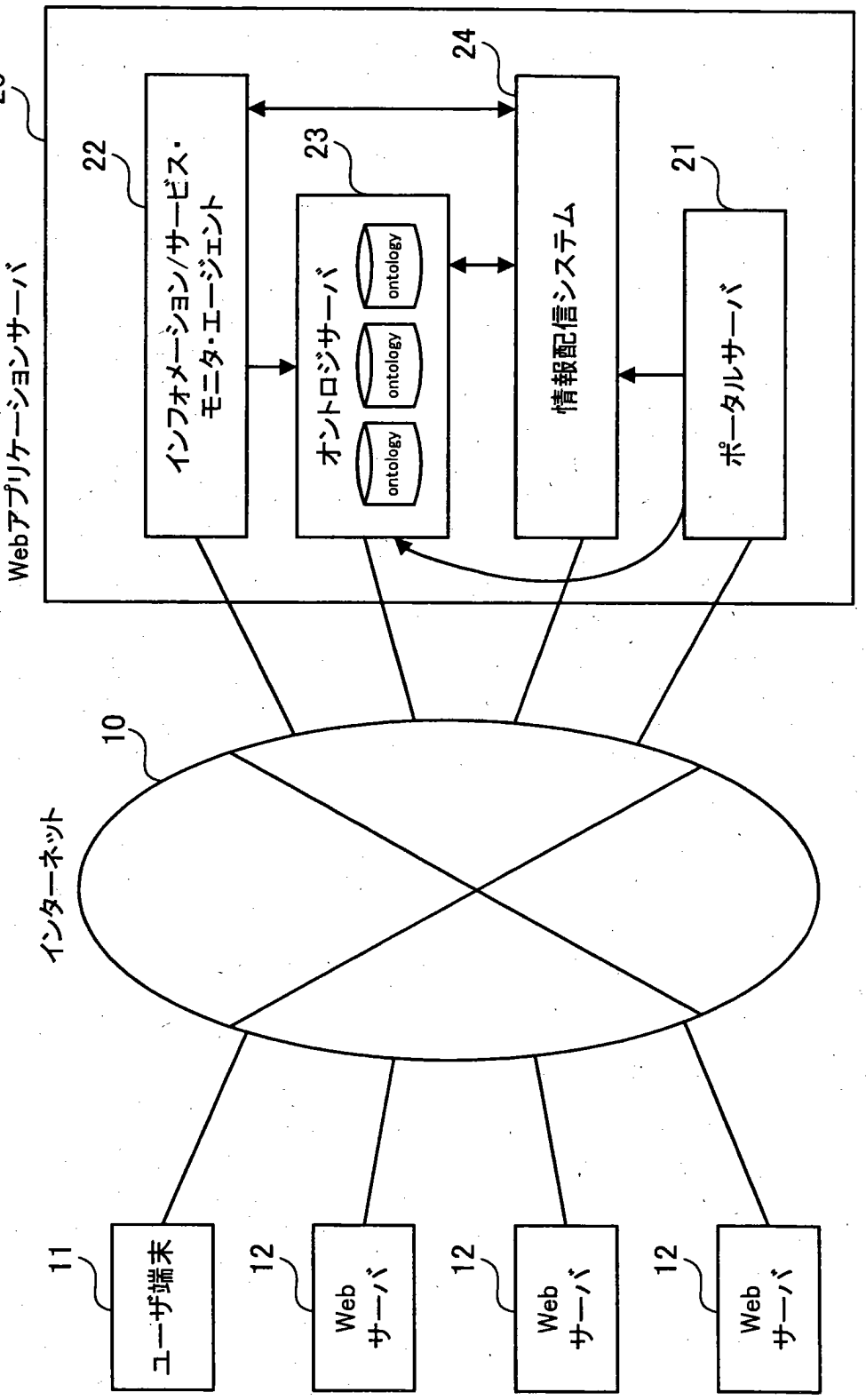
【図８】　本実施の形態における合算表示例を示した図である。

【符号の説明】

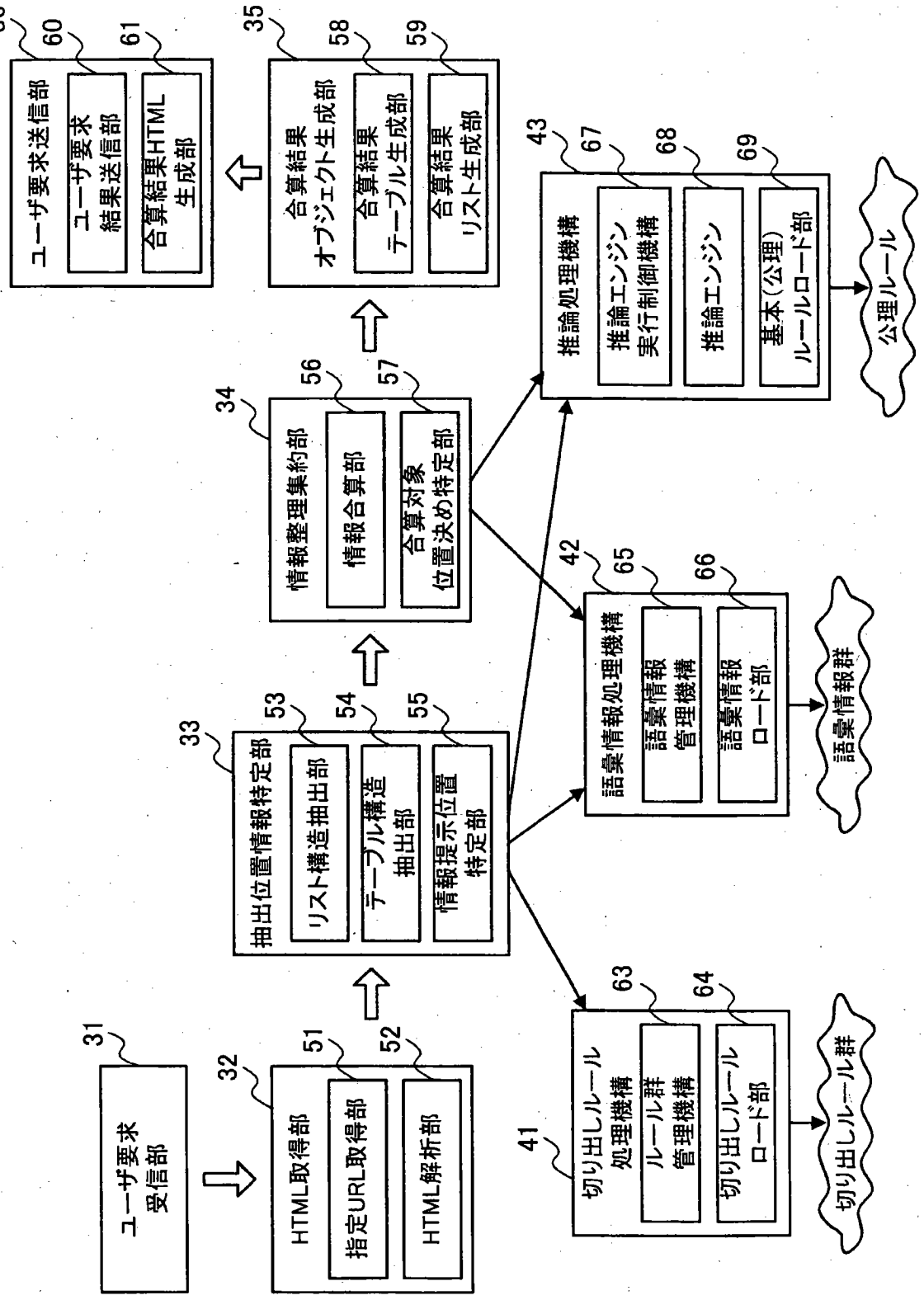
１０…インターネット、１１…ユーザ端末、１２…Ｗｅｂサーバ、２０…Ｗｅｂアプリケーションサーバ、２１…ポータルサーバ、２２…インフォメーション/サービス・モニタ・エージェント、２３…オントロジサーバ、２４…情報配信システム、３１…ユーザ要求受信部、３２…ＨＴＭＬ取得部、３３…抽出位置情報特定部、３４…情報整理集約部、３５…合算結果オブジェクト生成部、３６…ユーザ要求送信部、４１…切り出しルール処理機構、４２…語彙情報処理機構、４３…推論処理機構

【書類名】　　図面

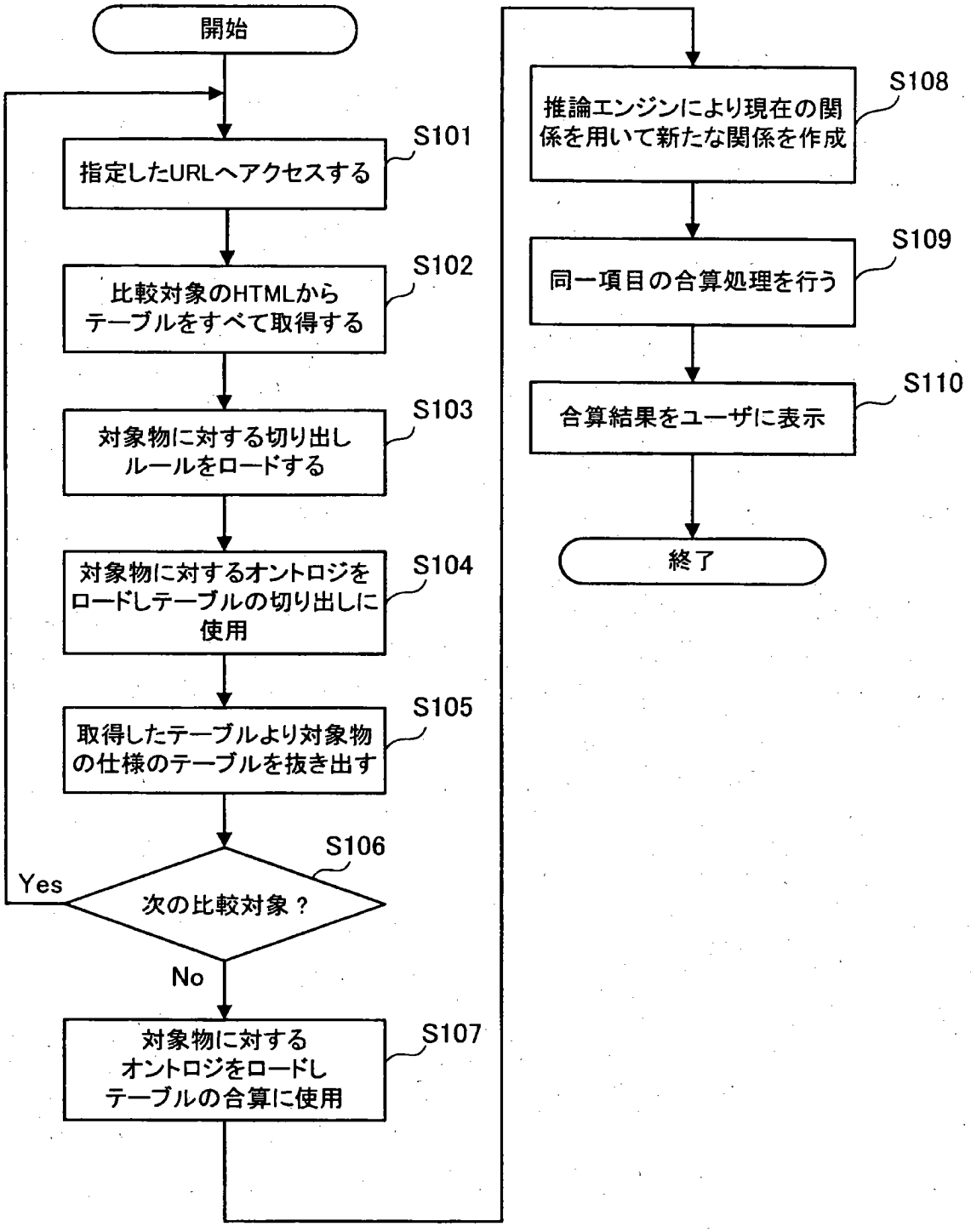
【図１】


20
Webアプリケーションサーバ
22
インフォメーション/サービス・
モニタ・エージェント
23
オントロジサーバ
ontology
ontology
ontology
24
情報配信システム
21
ポータルサーバ
10
インターネット
11
ユーザ端末
12
Web
サーバ
12
Web
サーバ
12
Web
サーバ

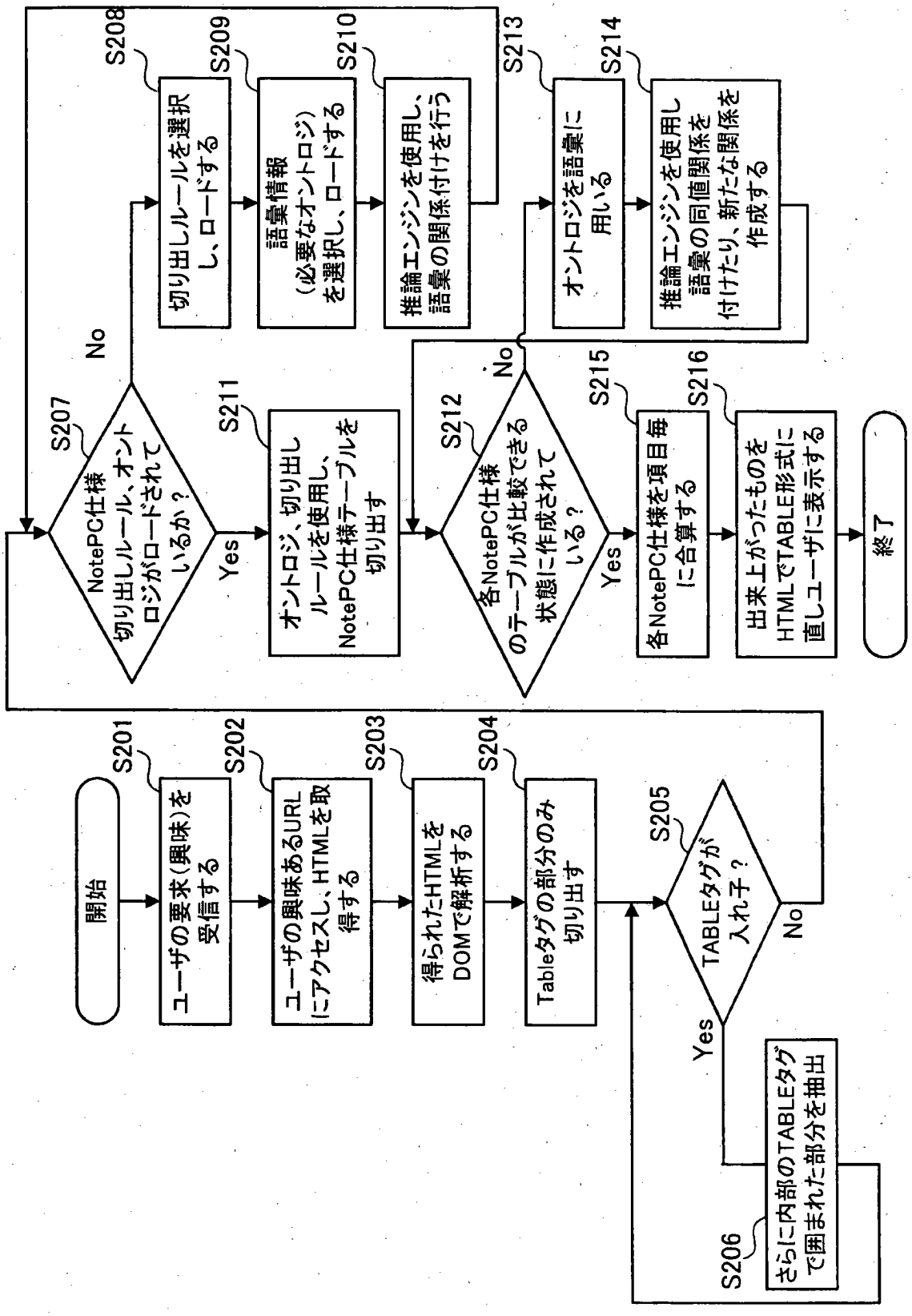
【図２】


ユーザ要求受信部
31
HTML取得部
32
指定URL取得部
51
HTML解析部
52
抽出位置情報特定部
33
リスト構造抽出部
53
テーブル構造抽出部
54
情報提示位置特定部
55
情報整理集約部
34
情報合算部
56
合算対象位置決め特定部
57
ユーザ要求送信部
36
ユーザ要求結果送信部
60
合算結果HTML生成部
61
合算結果オブジェクト生成部
35
合算結果テーブル生成部
58
合算結果リスト生成部
59
推論処理機構
43
推論エンジン実行制御機構
67
推論エンジン
68
基本(公理)ルールロード部
69
公理ルール
語彙情報処理機構
42
語彙情報管理機構
65
語彙情報ロード部
66
語彙情報群
切り出しルール処理機構
41
ルール群管理機構
63
切り出しルールロード部
64
切り出しルール群

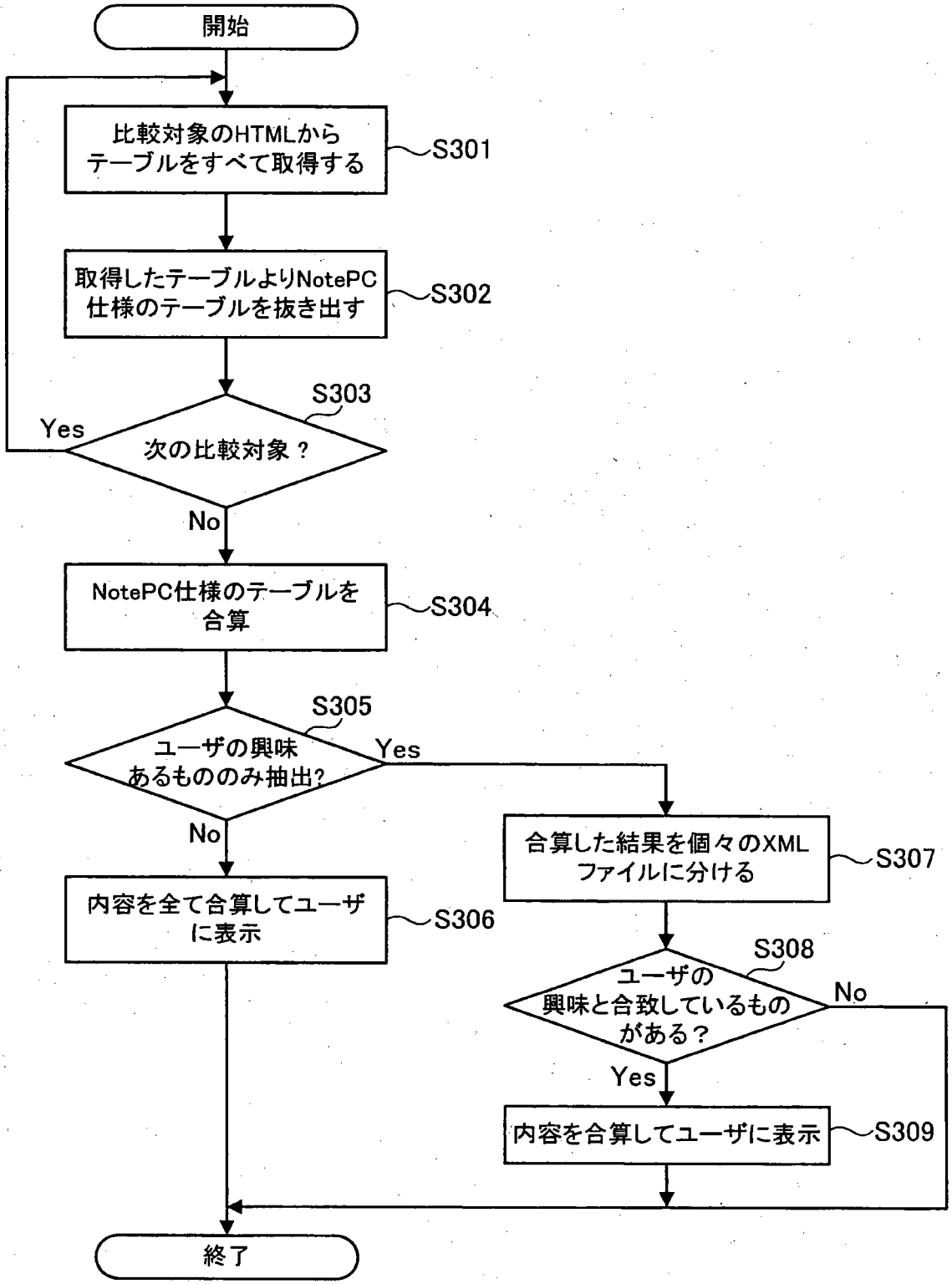
【図３】


開始
S101
指定したURLへアクセスする
S102
比較対象のHTMLから
テーブルをすべて取得する
S103
対象物に対する切り出し
ルールをロードする
S104
対象物に対するオントロジを
ロードしテーブルの切り出しに
使用
S105
取得したテーブルより対象物
の仕様のテーブルを抜き出す
S106
次の比較対象？
Yes
No
S107
対象物に対する
オントロジをロードし
テーブルの合算に使用
S108
推論エンジンにより現在の関
係を用いて新たな関係を作成
S109
同一項目の合算処理を行う
S110
合算結果をユーザに表示
終了

【図４】


開始
S201
ユーザの要求(興味)を受信する
S202
ユーザの興味あるURLにアクセスし、HTMLを取得する
S203
得られたHTMLをDOMで解析する
S204
Tableタグの部分のみ切り出す
S205
TABLEタグが入れ子？
Yes
No
S206
さらに内部のTABLEタグで囲まれた部分を抽出
S207
NotePC仕様 切り出しルール、オントロジがロードされているか？
No
Yes
S208
切り出しルールを選択し、ロードする
S209
語彙情報 (必要なオントロジ)を選択し、ロードする
S210
推論エンジンを使用し、語彙の関係付けを行う
S211
オントロジ、切り出しルールを使用し、NotePC仕様テーブルを切り出す
S212
各NotePC仕様のテーブルが比較できる状態に作成されている？
No
Yes
S213
オントロジを語彙に用いる
S214
推論エンジンを使用し語彙の同値関係を付けたり、新たな関係を作成する
S215
各NotePC仕様を項目毎に合算する
S216
出来上がったものをHTMLでTABLE形式に直しユーザに表示する
終了

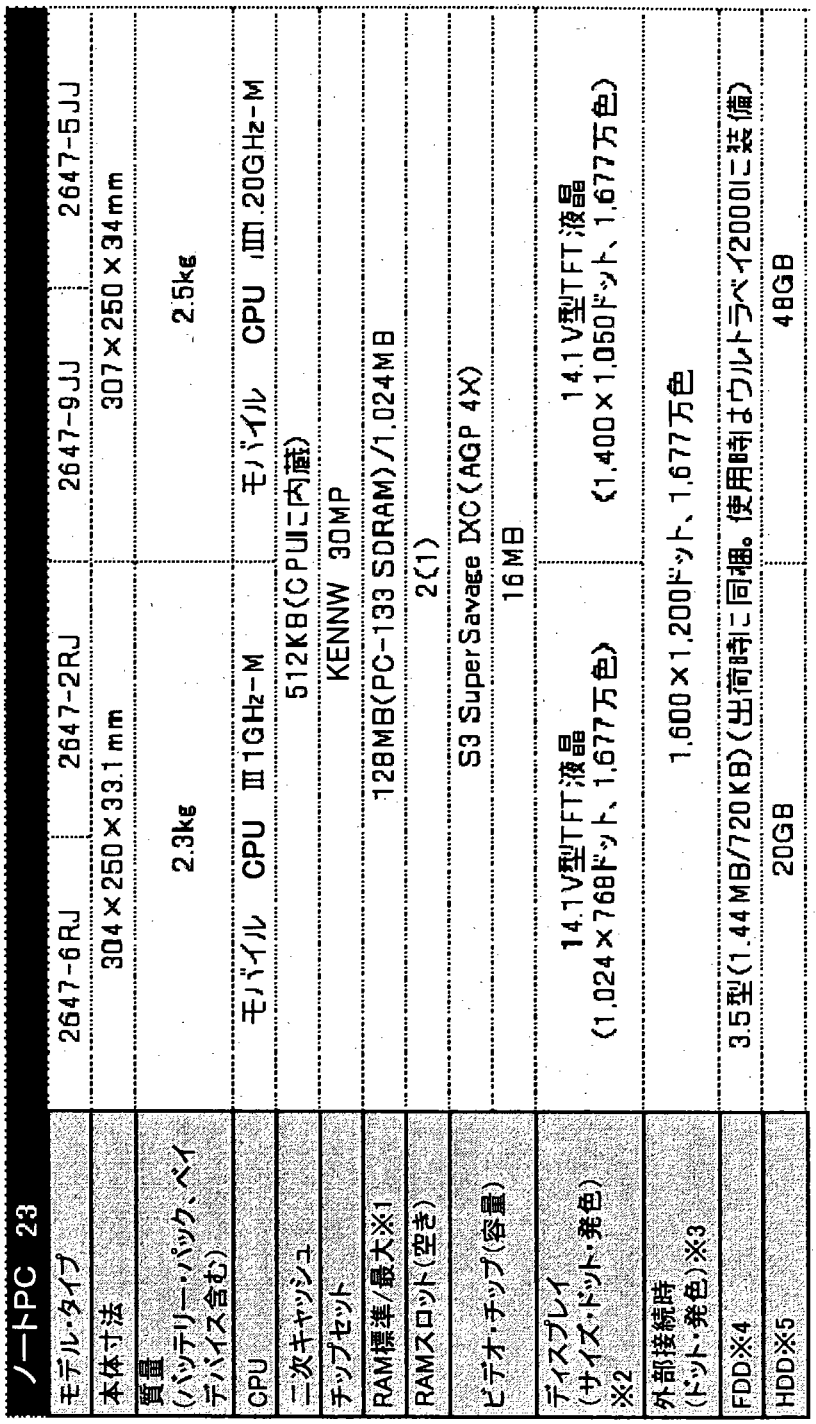
【図５】


開始
比較対象のHTMLから
テーブルをすべて取得する
S301
取得したテーブルよりNotePC
仕様のテーブルを抜き出す
S302
S303
次の比較対象？
Yes
No
NotePC仕様のテーブルを
合算
S304
S305
ユーザの興味
あるもののみ抽出?
Yes
No
内容を全て合算してユーザ
に表示
S306
合算した結果を個々のXML
ファイルに分ける
S307
S308
ユーザの
興味と合致しているもの
がある？
No
Yes
内容を合算してユーザに表示
S309
終了

【図６】

Order no.	M2638 J/F	M2699 J/F	M7802 J/F
プロセッサ	500MHz PPP PC G3	600MHz PPP PC：G3	600MHz PPP PC G3
二次キャッシュ	256KB/500MHz	256KB/600MHz	256KB/600MHz
メモリ（PC100 SDRAM）	128MB(640MBまで拡張可能)	128MB(640MBまで拡張可能)	256MB(640MBまで拡張可能)
システムバス	66MHz	100MHz	100MHz
ハードディスクドライブ（2）	約15GB Ultra ATA	約20GB Ultra ATA	約20GB Ultra ATA
光学式ドライブ	CD-ROM	DVD-ROM/CD-RWコンボ	DVD-ROM/CD-RWコンボ
ディスプレイ	12.1インチ（対角）TFT XGA	12.1インチ（対角）TFT XGA	14.1インチ（対角）TFT XGA
FireWire（IEEE 1394）	400Mbpsポート×1基		
USB	2基（各12Mbps）		
グラフィックスサポート	ATI RAGE Mobility 128（8MB SDRAM）		
VGAおよびコンポジットビデオ出力	VGAビデオ出力ポート（外部ディスプレイ／RGBプロジェクターとのビデオミラーリング用）		
コンポジットビデオ出力	コンポジットビデオ出力（AVポート経由でテレビ／AVプロジェクターに出力。別売のB社 AVケーブルが必要）		

【図７】



ノートPC 23				
モデル・タイプ	2647-6RJ	2647-2RJ	2647-9JJ	2647-5JJ
本体寸法	304×250×33.1mm		307×250×34mm	
質量（バッテリー・パック、ベイ・デバイス含む）	2.3kg		2.5kg	
CPU	モバイル CPU III1GHz-M		モバイル CPU III.20GHz-M	
二次キャッシュ	512KB（CPUに内蔵）			
チップセット	KENNW 30MP			
RAM標準/最大※1	128MB（PC-133 SDRAM）/1,024MB			
RAMスロット（空き）	2（1）			
ビデオ・チップ（容量）	S3 Super Savage IXC（AGP 4X） 16MB			
ディスプレイ（サイズ・ドット・発色）※2	14.1V型TFT液晶 （1,024×768ドット、1,677万色）		14.1V型TFT液晶 （1,400×1,050ドット、1,677万色）	
外部接続時（ドット・発色）※3	1,600×1,200ドット、1,677万色			
FDD※4	3.5型（1.44MB/720KB）（出荷時に同梱。使用時はウルトラベイ2000に装備）			
HDD※5	20GB		48GB	

【図８】

	A社ノートPC －仕様			[B社PC23 製品仕様]			
タイプ・モデル	M2638 J/F	M2699 J/F	M7802 J/F	2647-6RJ	2647-2RJ	2647-9JJ	2647-5JJ
プロセッサ	500MHz PPP PC G3	600MHz PPP PC G3	600MHz PPP PC G3	モバイル CPU 1GHz-M	モバイル CPU 1GHz-M	モバイル CPU 1.20GHz-M	モバイル CPU 1.20GHz-M
メモリ	128MB (640MBまで拡張可能)	128MB (640MBまで拡張可能)	256MB (640MBまで拡張可能)	128MB(PC-133 SDRAM)/1,024MB	128MB(PC-133 SDRAM)/1,024MB	128MB(PC-133 SDRAM)/1,024MB	128MB(PC-133 SDRAM)/1,024MB
ディスプレイ	12.1インチ(対角)TFT XGA	12.1インチ(対角)TFT XGA	14.1インチ(対角)TFT XGA	14.1V型TFT液晶 (1,024×768ドット、1,677万色)	14.1V型TFT液晶 (1,024×768ドット、1,677万色)	14.1V型TFT液晶 (1,400×1,050ドット、1,677万色)	14.1V型TFT液晶 (1,400×1,050ドット、1,677万色)
ハードディスクドライブ	A社 約15GB Ultra ATA	約20GB Ultra ATA	約20GB Ultra ATA	20GB	20GB	48GB	48GB
システムソフトウェア	A社 OS X version 10.1.2、Mac OS 9.2.2	A社 OS X version 10.1.2、A社 OS 9.2.2	A社 OS X version 10.1.2、A社 OS 9.2.2	XXX PX Professional	XXX 2002 Professional	XXX PX Professional	XXX 2002 Professional

【書類名】　　要約書

【要約】

【課題】　例えばＷｅｂ上に公開されている様々な領域のカタログ等を、自動的に切り出す。

【解決手段】　ユーザの興味に関する情報を受信するユーザ要求受信部３１と、受信した情報に基づいて、複数のサイトからＨＴＭＬ文書を取得するＨＴＭＬ取得部３２と、取得したＨＴＭＬ文書に対して切り出し処理を施すための切り出しルールを提供する切り出しルール処理機構４１と、受信した情報に基づいてオントロジを読み出し、語彙情報を得る語彙情報処理機構４２と、公理ルールに基づいて推論演算を実行する推論処理機構４３と、取得したＨＴＭＬ文書に対し、切り出しルール処理機構４１の切り出しルール、語彙情報処理機構４２からの語彙情報、推論処理機構４３の推論演算に基づき、ＨＴＭＬ文書のタグを頼りに抽出データオブジェクトを取り出す抽出位置情報特定部３３とを含む。

【選択図】　　図２

認定・付加情報

特許出願の番号	特願２００２－２１８７４０
受付番号	５０２０１１０８８５１
書類名	特許願
担当官	佐々木　吉正　　２４２４
作成日	平成１４年１１月２９日

＜認定情報・付加情報＞

【特許出願人】

【識別番号】　390009531

【住所又は居所】　アメリカ合衆国１０５０４、ニューヨーク州　アーモンク　ニュー　オーチャード　ロード

【氏名又は名称】　インターナショナル・ビジネス・マシーンズ・コーポレーション

【代理人】

【識別番号】　100086243

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名又は名称】　坂口　博

【代理人】

【識別番号】　100091568

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名又は名称】　市位　嘉宏

【代理人】

【識別番号】　100108501

【住所又は居所】　神奈川県大和市下鶴間１６２３番１４　日本アイ・ビー・エム株式会社　知的所有権

【氏名又は名称】　上野　剛史

【復代理人】　申請人

【識別番号】　100104880

【住所又は居所】　東京都港区赤坂５－４－１１　山口建設第２ビル６Ｆ　セリオ国際特許事務所

【氏名又は名称】　古部　次郎

次頁無

【書類名】　出願人名義変更届

【あて先】　特許庁長官殿

【事件の表示】

【出願番号】　特願2002-218740

【承継人】

【識別番号】　592073101

【氏名又は名称】　日本アイ・ビー・エム株式会社

【承継人代理人】

【識別番号】　100086243

【弁理士】

【氏名又は名称】　坂口　博

【連絡先】　０４６－２１５－３３２５

【承継人代理人】

【識別番号】　100091568

【弁理士】

【氏名又は名称】　市位　嘉宏

【連絡先】　０４６－２１５－３３２５

【承継人代理人】

【識別番号】　100108501

【弁理士】

【氏名又は名称】　上野　剛史

【連絡先】　０４６－２１５－３３２５

【手数料の表示】

【予納台帳番号】　029193

【納付金額】　4,200円

【提出物件の目録】

【包括委任状番号】　0004471

【包括委任状番号】　0004470

【包括委任状番号】　0208086

【プルーフの要否】　　要

認定・付加情報

特許出願の番号	特願２００２－２１８７４０
受付番号	５０２０１５５３７８０
書類名	出願人名義変更届
担当官	佐々木　吉正　　２４２４
作成日	平成１５年　１月　９日

＜認定情報・付加情報＞

【提出日】　平成14年10月16日

【承継人】

【識別番号】　592073101

【住所又は居所】　東京都港区六本木３丁目２番１２号

【氏名又は名称】　日本アイ・ビー・エム株式会社

【承継人代理人】　申請人

【識別番号】　100086243

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名又は名称】　坂口　博

【承継人代理人】

【識別番号】　100091568

【住所又は居所】　神奈川県大和市下鶴間１６２３番地１４　日本アイ・ビー・エム株式会社　大和事業所内

【氏名又は名称】　市位　嘉宏

【承継人代理人】

【識別番号】　100108501

【住所又は居所】　神奈川県大和市下鶴間１６２３番１４　日本アイ・ビー・エム株式会社　知的所有権

【氏名又は名称】　上野　剛史

次頁無

出　願　人　履　歴　情　報

識別番号　〔３９０００９５３１〕

１．変更年月日　２００２年　６月　３日
［変更理由］　住所変更
住　所　アメリカ合衆国１０５０４、ニューヨーク州　アーモンク　ニュー　オーチャード　ロード
氏　名　インターナショナル・ビジネス・マシーンズ・コーポレーション

出願人履歴情報

識別番号　〔５９２０７３１０１〕

１．変更年月日　１９９２年　４月　３日
［変更理由］　新規登録
住　所　東京都港区六本木３丁目２番１２号
氏　名　日本アイ・ビー・エム株式会社